人文研究与电子信息技术

⊙ 张隆溪

科技与人文似乎是人类智力活动的两个极端，但电子信息技术的发展，已经出现了所谓“数码化人文研究”(digital humanities) 这一新名词和新领域。早在计算机和数码技术出现之前，语言学和人文研究就已经开始使用数据统计的方法。这种方法一个基本的应用就是书籍的引得，index，现在一般译为索引。西方出版的学术著作往往有详略程度不等的参考书目和索引，对学术研究很有帮助。1930 年哈佛燕京学社成立引得编纂处，以洪业主持其事，从二十世纪三十年代开始，前后出版了六十多种引得，对于海外的汉学研究起了很大作用，对中国学者研究自己的古籍也非常有用。诚如陈毓贤女士在其所著《洪业传》里所说，“这些引得在中国研究古籍的学术上创立了新纪元”[1]。在传统学术研究中，虽然博闻强记非常重要，却没有以具体字句的统计为基础的科学方法，而由西方引进的索引，就开创了一种全新的研究方法和研究工具，使学术研究得以离开印象式的评论而具有更扎实的文本基础。到电子数码技术出现之后，很快就有了更精确而且大型的引得如 *Shakespeare Concordance* 之类的参考书。香港中文大学中国古籍研究中心出版的先秦两汉古籍逐字索引和魏晋南北朝古籍逐字索引，在中国古代典籍索引方面，也是非常重要的参考书。这些引得或索引对于检索某个字句的来源出处非常方便，也可以帮助我们了解各个词汇在不同时代、不同典籍里相关但有时不尽相同的用法。

数据统计方法的另一个应用就是语言词汇出现频率的统计。美国教育心理学家桑戴克 (Edward L. Thorndike，1874—1949) 以词汇出现频率编制最常用词汇表，语言学家高名凯先生就依据桑戴克的理论，编过一部《英语常用词汇》，1961 年由北京商务印书馆出版。学习一种语言，掌握常用词汇相当重要，而常用词的定义就是以词汇使用频率为科学的依据。词汇频率和索引结合起来，往往可以产生在具体材料上有确切依据的观点和看法，对于学术研究产生很大影响。具体就文学研究而言，一个作家或诗人使用词汇的频率，往往可以告诉我们这位作家或诗人文体风格的特点，并且由此引导我们窥见其思想情感的隐秘，在文学批评和研究方面有很大帮助。钱钟书先生在《谈艺录》里就曾用数节文字来讨论唐代诗人李贺诗中用字，他虽然没有严格统计用字的频率，但依据长吉诗中用字多少来立论，其实和计算词汇频率是同一个道理。钱钟书说法国诗人“戈蒂埃(Gautier)作诗文，好镂金刻玉。其谈艺篇(L’Art)亦谓诗如宝石精镠，坚不受刃(le bloc résistant)乃佳，故当时人有至宝丹之讥 (le matérialisme du style)。…… 近人论赫贝尔(F. Hebbel)之歌词、爱伦坡(E. A. Poe)之文、波德莱尔(Baudelaire)之诗，各谓三子好取金石硬性物作比喻。…… 窃以为求之吾国古作者，则长吉或其伦乎”。接下去他引了李贺诗中许多例证：“如《李凭箜篌引》之‘昆山玉碎凤凰叫’，‘石破天惊逗秋雨’；《残丝曲》之‘缥粉壶中沉琥珀’；《梦天》之‘玉轮轧露湿团光’；《唐儿歌》之‘头玉硗硗眉刷翠’；《南园》之‘晓月当帘挂玉弓’；《十二月乐词》之‘香汗沾宝粟，夜天如玉砌’；《秦王饮酒》之‘羲和敲日玻璃声’；《马诗》之‘向前敲瘦骨，犹自带铜声’；《勉爱行》之‘荒沟

古水光如刀’;《春归昌谷》之‘谁揭赪玉盘,东方发红照’;《江南弄》之‘酒中倒卧南山绿,江上团团贴寒玉’;《北中寒》之‘山湿无声玉虹寒’;《溪晚凉》之‘玉烟青湿白如幢’;《将进酒》之‘琥珀浓,小槽酒滴珍珠红’等等”。后面钱钟书还说,李贺诗里常用“凝”字,“至其用‘骨’字、‘死’字、‘寒’字、‘冷’字句,多不胜举,而作用适与‘凝’字相通”[2]。再后面他又说李贺诗中“好用青白紫红等颜色字”,那是一般读者都容易注意到的现象,钱钟书则认为“尚是描画皮毛,非命脉所在也”[3]。可见通过李贺诗中常用的金、石、玉、琥珀等字,可以概括出他的作品语言给读者一种冷峻、刚硬的感觉,其中有许多诗句想象奇巧叠出,如《秦王饮酒》之“羲和敲日玻璃声”,《马诗》之“向前敲瘦骨,犹自带铜声”,用“敲”字带出玻璃和金属的声音,给人裂冰碎玉那种硬而脆的感觉,就造成李贺诗特有风格的印象,使我们意识到李贺的确像戈蒂埃等欧美诗人一样,“好取金石硬性物作比喻”,在风格上可以相比。这一概括以具体词语的使用频率为基础,就很有说服力。如果我们现在用电子信息技术对李贺的文本做一个词汇频率统计,就更能够证明这一点。

在西方文学研究中,卡洛淋·斯佩琴(Caroline Spurgeon)在1935年发表了《莎士比亚的意象及其意义》(*Shakespeare's Imagery and What It Tells Us*)一书,就是以统计的方法来研究莎士比亚的文体风格,并由此探讨作家的思想。这在当时的文学研究中很有新意,曾引起一阵轰动,而且这本书历年来一直重印再版。虽然斯佩琴主要研究《坎特伯雷故事集》(*Canterbury Tales*)的作者乔叟(Geoffrey Chaucer,1343—1400),她现在还被人记得的却是这部研究莎士比亚意象的书,在六十多年之后,剑桥大学出版社于1993年还重印了此书。这本书除了一般学术著作的文字叙述之外,在书后有六个图表,把莎士比亚作品中一些重要意象使用频率用图表标示出来,还把他同时代几位作家使用的意象绘成图表,以比较他们的异同。斯佩琴认为,有时候一个突出的意象贯穿莎剧整部作品,例如《李尔王》全剧都给人挣扎、痛苦,甚至一种肉体的、肌肤煎熬之痛的感觉,而这一感觉就来自剧中不断使用身体受苦的动词和相关意象。“只要打开这个剧本任何一页,都很难不被这些意象和动词所震撼,因为每一种身体的动作,往往是痛苦的身体动作,都用来表现不止于实际上肉体的疼痛,而且也表现精神和抽象的痛苦”[4]。许多人读《李尔王》等莎士比亚作品,都会形成一定的印象,但这种印象是笼统而不明确的,斯佩琴用统计方法把这些印象落实到具体的意象和词汇,就使印象式批评有了具体文本的依据,造成一种类似科学式的批评,在文学研究中独辟蹊径,很有影响。虽然后来的研究者们大都没有严格计算意象频率,但注重意象成为文学研究中一个十分重要的方法。这就是说,对具体词汇和意象的把握是讨论文学作品一个非常重要的方面,而在这方面,现代飞速发展的信息科技就可以为文学研究提供更多更便利的研究工具。

十年前电子版《四库全书》的出版,可以说是数码化技术一个新的里程碑。这项大工程把文渊阁四库全书全部数码化,使之成为可以搜索的电子文本。美国加州大学圣塔芭芭拉分校艾朗诺教授曾撰文详细谈论他使用电子版《四库全书》的经验,认为“就其极大规模和多种用途而言,这电子版《四库全书》势必在我们古代中国研究这一人文领域的研究方法上,留下它的印迹”[5]。首先是其规模,其次是其搜寻速度,两者加在一起,就可以在过去不可能想象的速度和范围内,搜寻具体的词句和意象,而且可以根据搜索的结果,看出某个主题或关键词在各种书籍里出现的频率。艾朗诺教授是研究宋代文学的知名学者,他举例说明在他的研究中,电子版《四库全书》如何给他提供许多帮助。例如他通过查询电子版《四库全书》,对瓷器在宋诗中出现的情形就得出了可靠而出乎意料的结论。虽然宋瓷非常雅致精美,在宋人生活中应该是日常所用所见,也应该是宋代文人所把玩的,可是通过查询电子版《四库全书》,艾朗诺很快发现宋诗里很少写到瓷器,只是偶尔提到茶碗。他说:“没有电子查询,我绝不可能察觉到瓷器在宋诗里这出乎意料的分布情形,即有时候提到,却并没有特别注意。同样重要的是,没有电子数据库的依据,就须花费数月甚至数年时间的阅读,才可能对自己作出的结论产生信心,而有电子搜寻技术的帮助,就很快可以做到这一点。有了电子版《四库全书》的帮助,我们就可以减少我们结论当中凭印象得来的方面,而更接近在统计上有依据的、可

以客观验证的结果。”[6]对于人文学者说来，电子信息技术的应用使人文研究在某些方面减少了随意性质，而似乎更近于科学，这是人文学者非常重视的一点。

不过艾朗诺教授也指出一些在使用电子版《四库全书》当中发现的问题，最主要是搜索范围的问题。以关键词为单位得出的结果往往太多，无法对一个具体词语出现的环境做进一步限定，于是出现成百上千太多的“匹配”而变得没有什么用处[7]。对于文学研究或广义的人文研究而言，确定语言词汇的意义在研究中具有核心作用，而词汇的意义在任何一个文本（text）中，都取决于上下文的语境（context），于是语境非常重要，而语境总是具体的，对意义的确定具有限定作用。电子文本的好处是可以快速搜索任何词语，但怎样使搜索范围接近具体的、对一个具体词语的意义有限定作用的语境和范围，则是现在仍然有待解决的问题。就以编制书籍索引为例，简单的索引只列出书中提到的人名或其它关键词语，那种索引大概比较容易用计算机完成，但那种索引对读者并没有很大帮助。更详尽的多层次索引才更有帮助，但那就涉及对关键词语及其相互关联的判断，是作者本人最能够知道和制作的，用机械的方法就很难做这样的索引。

与此相关的机器翻译问题，也可以说明这一点。到目前为止，机器翻译较能应付的是比较简单、程序化的语句，而比较复杂的文本，尤其是文学作品，就不可能用机器来翻译。我想这其中原因，就是读者可以以自己的语言能力、阅读经验和文学常识为基础，判断在一个上下文的具体语境里，一个词语和意象具有什么意义，应该如何理解。这当中很难归纳出一条普遍适用的规律，也就很难设计出可以机械操作的程序，再转换成计算机的语言。文学语言往往不是只有一种理解，一种解释，尤其是诗的语言，往往利用意义的含蓄多义，造成多种理解的可能，那正是文学丰富意蕴之所在。从这个意义上说，文学和科学之间，大概总存在一种紧张的关系。作为使用工具的动物，可以说人从一开始就在制造各种器具来代替人工，尤其在近代大工业机械化生产出现以来，机器的作用越来越大，从工业、农业、军事、建筑、旅行、通讯直到我们的日常生活，机器取代人工都成为普遍趋势。尤其进入二十一世纪以来，电子技术的发展和所谓数码革命，使机器取代人本身已经在人们的想象中以各种形式呈现。恰恰在文学的领域，各种科幻小说已经想象 cyborg 的出现，打破了人和机械的界限，幻想各种比人更有智慧、更强壮、有更高文明程度的机器人。这种幻想与现实之间的关系，也随着技术的发展而缩小。

然而一般说来，人文学者对机器（包括现在越来越发展的计算机和数码技术）能够达到甚至超越人脑的综合分析能力，总是抱着怀疑态度。在人的身体活动能力方面，机器的确往往可以代替人，而且比人更迅速有效。人没有虎豹的獠牙利爪，却可以制造武器比任何动物都更具杀伤力；人没有鹞鹰的羽翼，却可以制造飞机比任何禽鸟飞得更高更快，环球旅行。我们日常生活中有各种机器帮助我们做各种事情，那往往是靠人的体力很难、甚至根本无法完成的任务。然而在人的智力和思想方面，在想象和审美经验方面，我们却总相信人是不可以取代的。因此，一个人文学者希望于科学技术的是研究工具和研究方法的更加丰富完善，但不是取代人的脑力劳动，也不可能取代人的现实感和想象。有人认为有了计算机，有了数据库，过去强调那种博闻强记已经毫无意义，这还是言之过早，而且是言过其实。尤其就人文研究而言，记忆不是机械的，而是在思考问题时可以产生联想、见出事物之间联系的基础，而计算机和机器不可能轻易取代。至于计算机信息技术究竟能为我们提供什么，则有待科技专家们为人文学者展示我们还远远不知道、不了解的科学的奇观。

注释：

〔1〕陈毓贤，《洪业传》，台北：联经 1992 年版，第 171 页。

〔2〕〔3〕钱钟书：《谈艺录（补订本）》，中华书局 1984 年版，第 48—49、51 页。

〔4〕Caroline Spurgeon, Shakespeare's Imagery and What It Tells Us (Cambridge: Cambridge University Press, 1990[1935]), p. 339.

〔5〕〔6〕〔7〕Ronald Egan, "Reflections on Uses of the Electronic Siku quanshu," Chinese Literature: Essays, Articles, Reviews 23 (2001): 103、107、111—112.